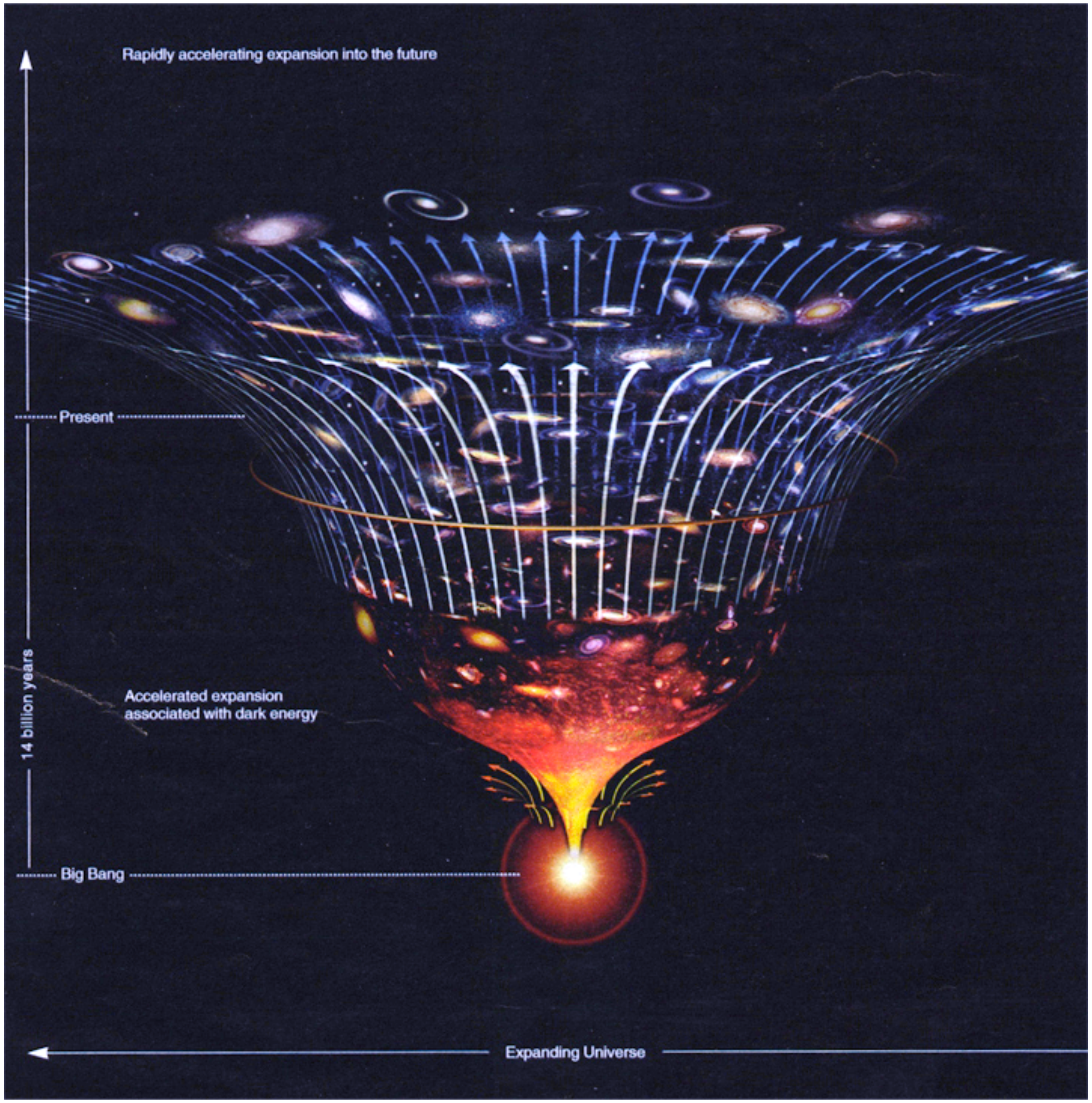
Spectroscopic Databases and Manifold Learning for Surveys of the 2020s

Dan Masters

Jet Propulsion Laboratory, California Institute of Technology

December 6, 2018

Jet Propulsion Laboratory
California Institute of Technology



Breakthrough from the 1990s:
Accelerating cosmic expansion

2011 Nobel Prize in Physics

Λ?

The Redshift Measurement Problem

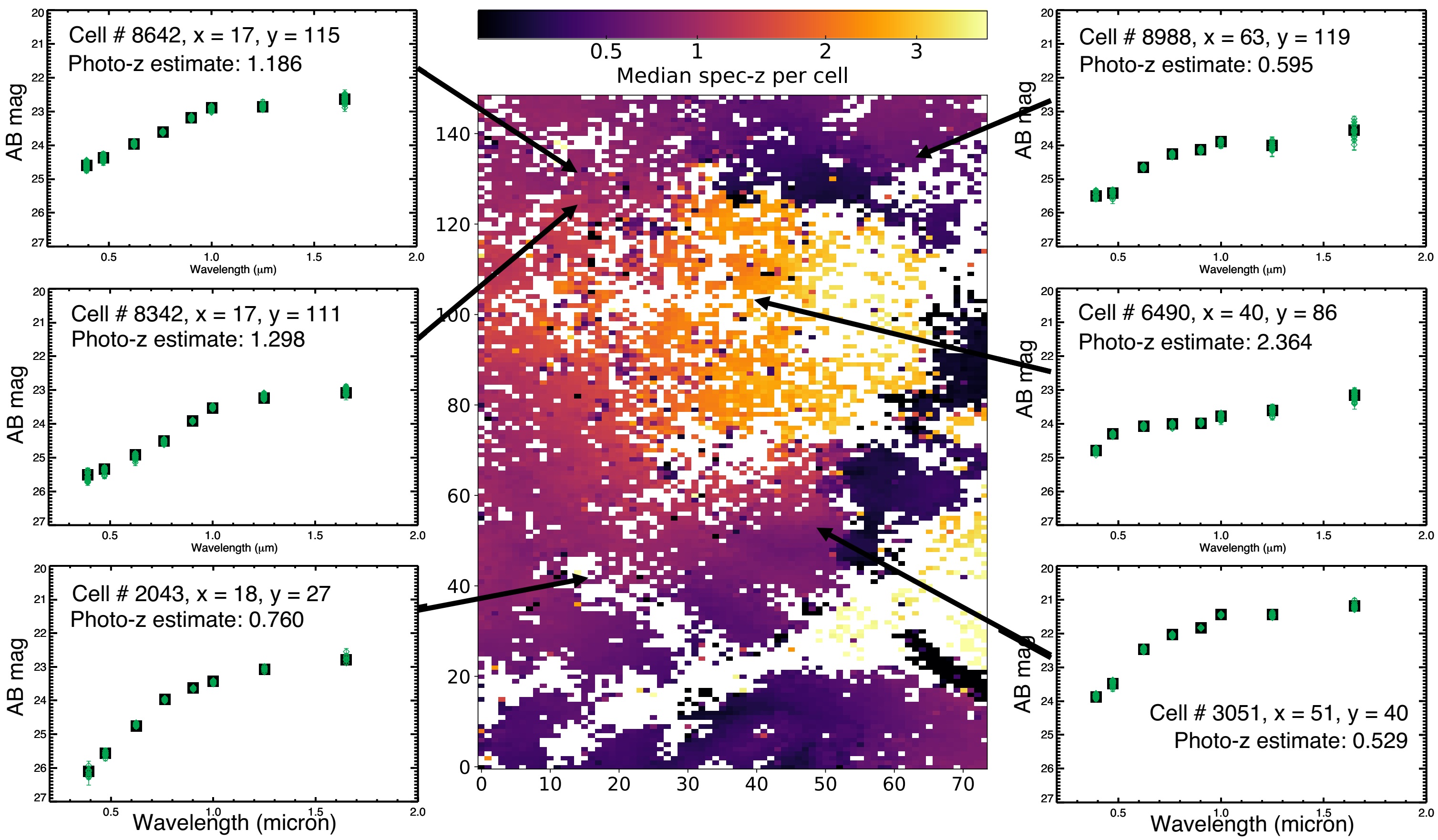
- Billions of galaxies will be imaged by the Stage IV cosmology surveys (LSST, Euclid, WFIRST)
- Only possible to get spectroscopic redshifts for a small fraction
- Weak lensing cosmology requires that the redshift distributions of galaxies in ~10-20 redshifts bins be known with **high accuracy**

→**Photometric redshifts will necessarily be crucial for weak lensing cosmology missions**

Manifold learning / nonlinear dimensionality reduction (NLDR)

- Group of techniques to characterize / explore high-dimensional data and correlations in high dimensions
- Common ones includes the self-organizing map (SOM), t-SNE, local linear embedding (LLE), and UMap
- Most project the high-D manifold down to a lower-D representation
- Whereas deep convolution networks try to learn a complex high-dimensional relationship between input data and output labels, NLDR just tries to unwrap the high-D data in an unsupervised way – no outputs

Self-organized map of galaxy colors to Euclid depth



Masters et al. 2015, ApJ, 813, 53

The galaxy color manifold

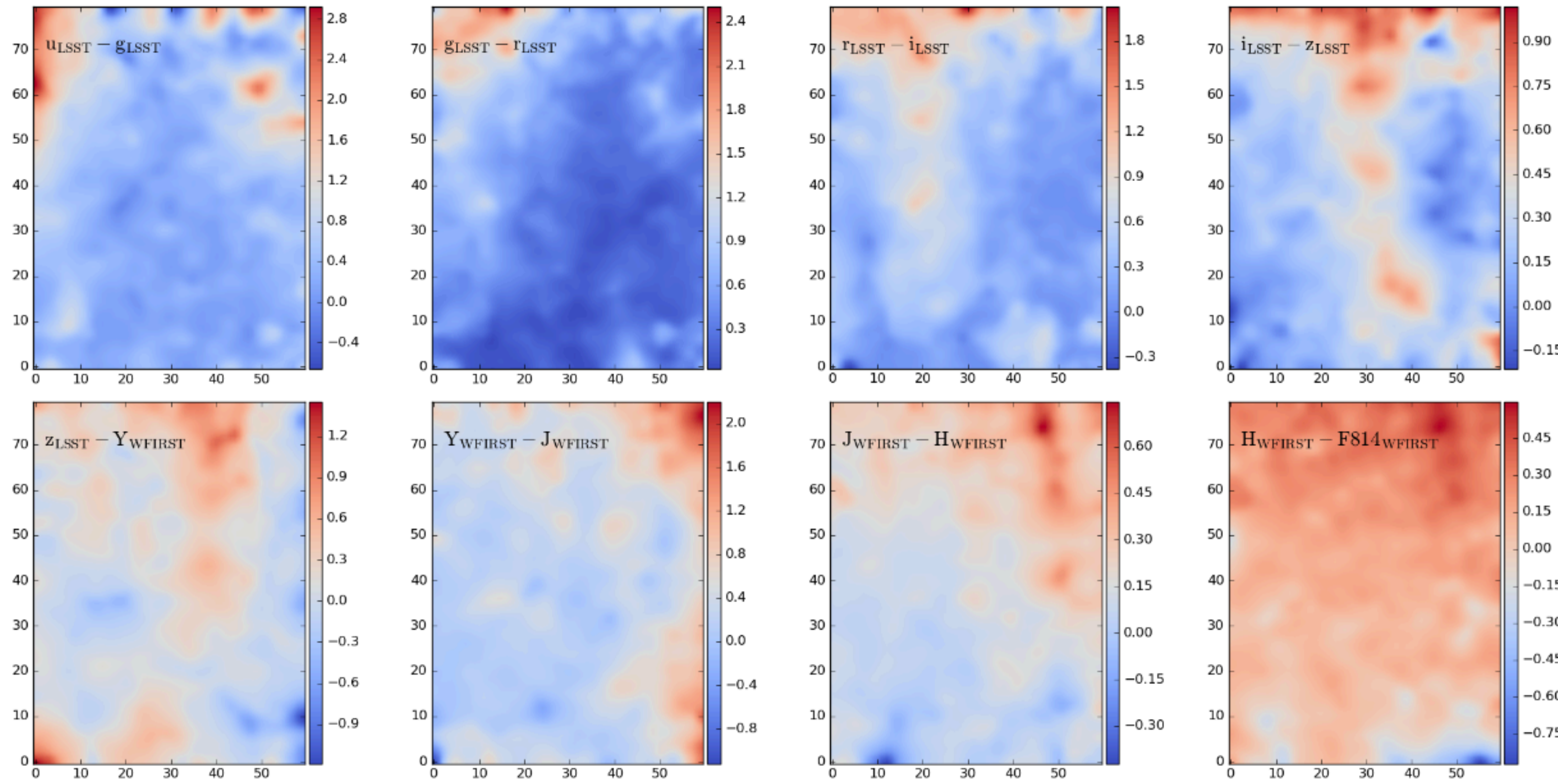
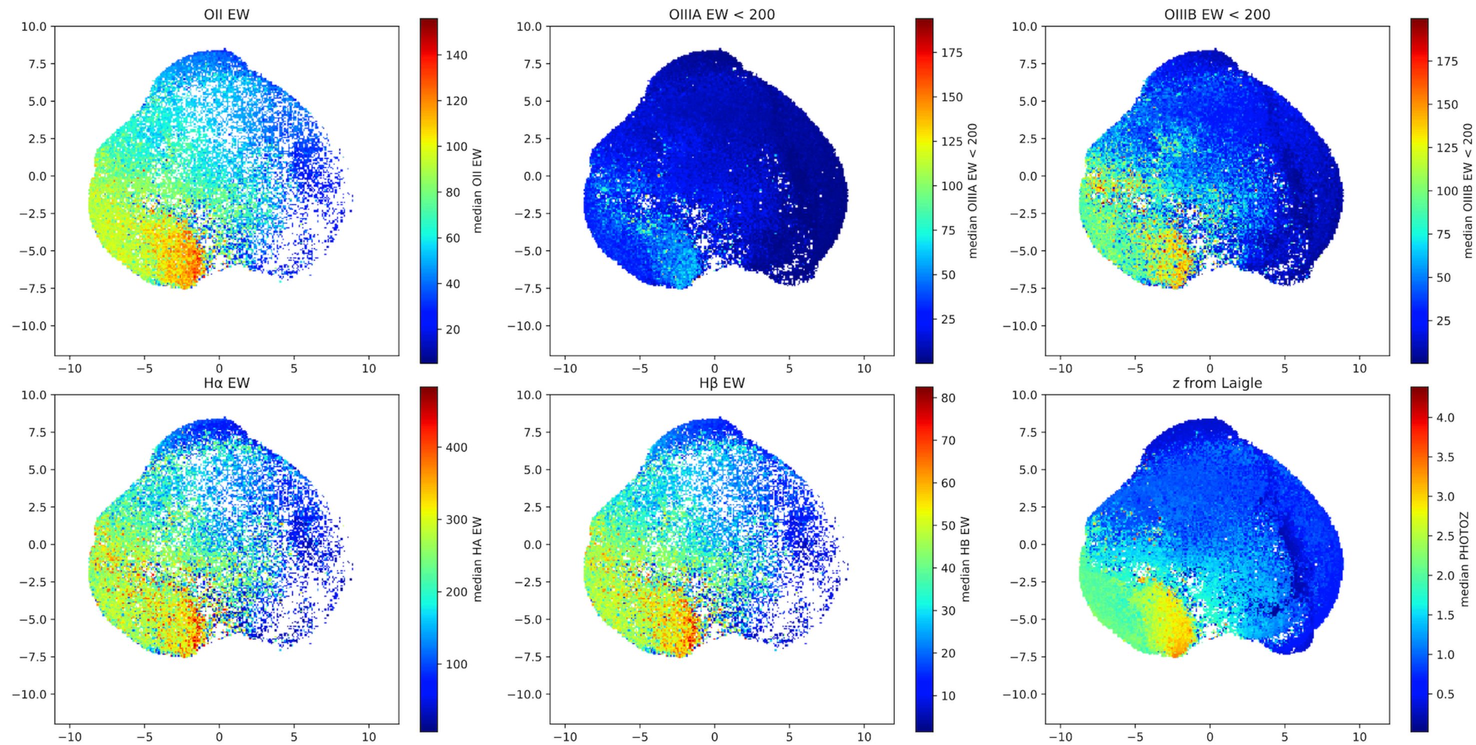


Figure 5. LSST and *WFIRST* colors of the trained SOM at each cell from top-left to bottom-right color-coded by: $\mathrm{u_{LSST} - g_{LSST}}$, $\mathrm{g_{LSST} - r_{LSST}}$, $\mathrm{r_{LSST} - i_{LSST}}$, $\mathrm{i_{LSST} - z_{LSST}}$, $\mathrm{z_{LSST} - y_{WFIRST}}$, $\mathrm{y_{WFIRST} - j_{WFIRST}}$, $\mathrm{j_{WFIRST} - H_{WFIRST}}$, and $\mathrm{H_{WFIRST} - F184W_{WFIRST}}$. SOM is selected to be a mesh of 80×60 cells. The axes are arbitrary and each position on the two dimensional map points to a position in the 8 dimensional color space.

Hemmati et al. 2018

Other techniques – UMap



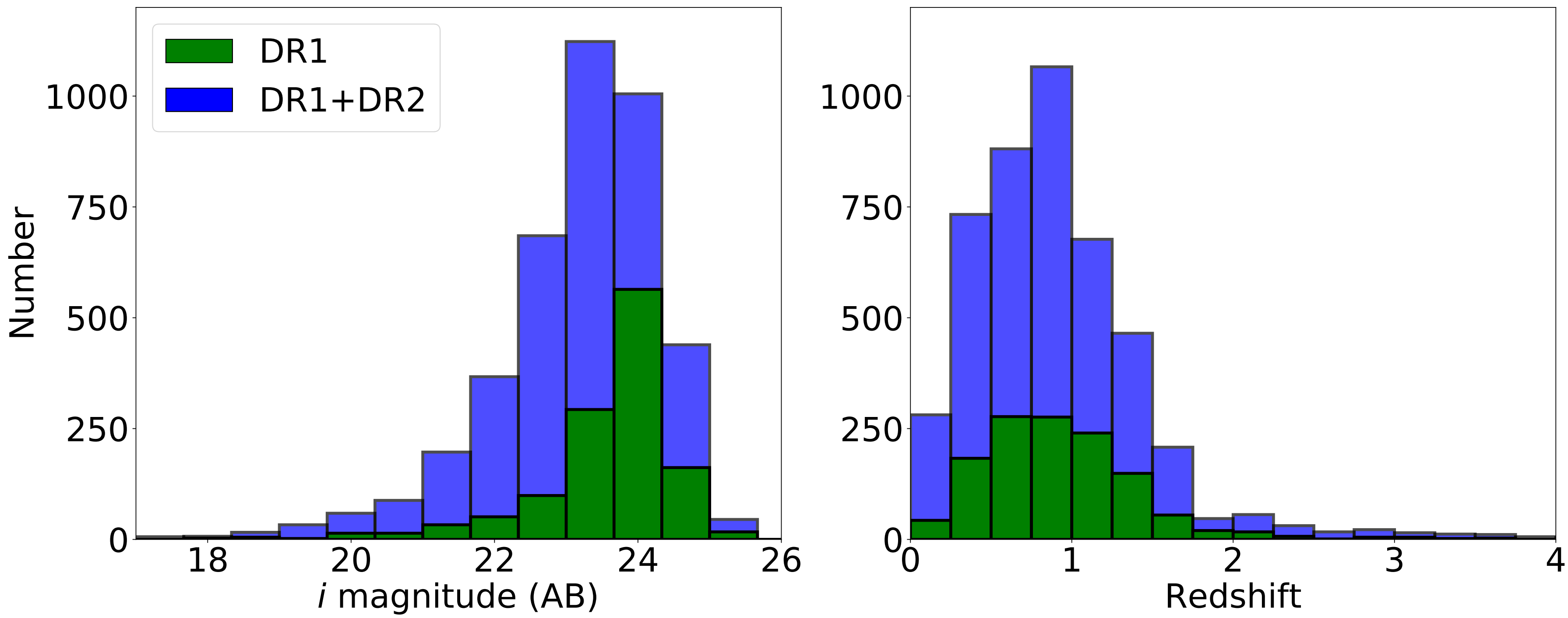
C3R2 = Complete Calibration of the Color-Redshift Relation

Judith Cohen (Caltech) - PI of Caltech Keck C3R2 allocation
16 nights (DEIMOS + LRIS + MOSFIRE, kicked off program in 2016A)
Daniel Stern (JPL) - PI of NASA Keck C3R2 allocation
10 nights (all DEIMOS; "Key Strategic Mission Support")
Daniel Masters (JPL) – PI of NASA Keck C3R2 allocation 2018A/B. **Observed last night, and will again tonight**
10 nights (5 each LRIS/MOSFIRE; "Key Strategic Mission Support")
Dave Sanders (IfA) - PI of Univ. of Hawaii Keck C3R2 allocation
6 nights (all DEIMOS) + H20
Bahram Mobasher (UC-Riverside) - PI of UC Keck C3R2 allocation
2.5 nights (all DEIMOS)

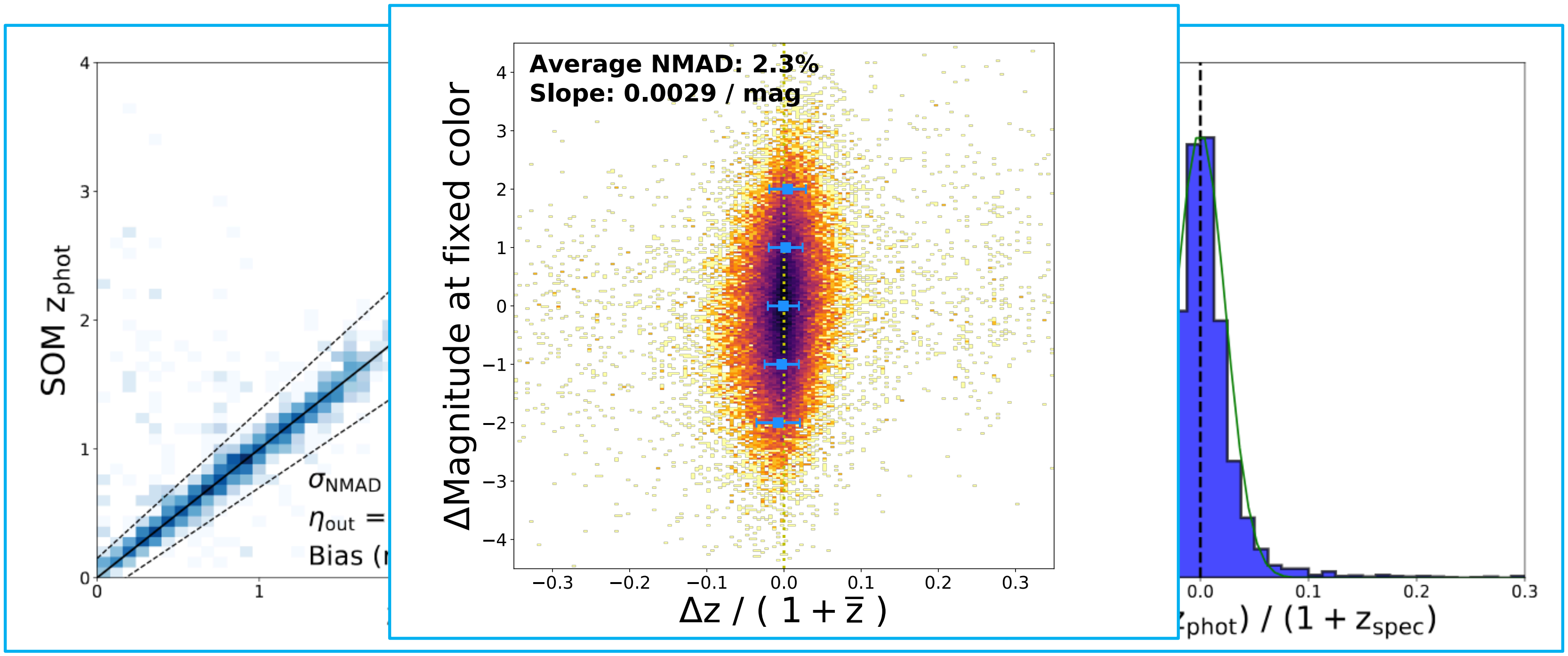
+ time allocations on VLT (PI F. Castander), MMT (PI D. Eisenstein), and GTC (PI C. Guitierrez)
-Sample drawn from 6 fields totaling ~6 deg^2

Additional Collaborators: Peter Capak, S. Adam Stanford, Nina Hernitschek, Francisco Castander, Sotiria Fotopoulou, Audrey Galametz, Iary Davidzon, Stephane Paltani, Jason Rhodes, Alessandro Rettura, Istvan Szapudi, and the Euclid Organization Unit – Photometric Redshifts (OU-PHZ) team

C3R2-Keck stats through DR2 (2016A-2017A)



C3R2 – Results from SOM method

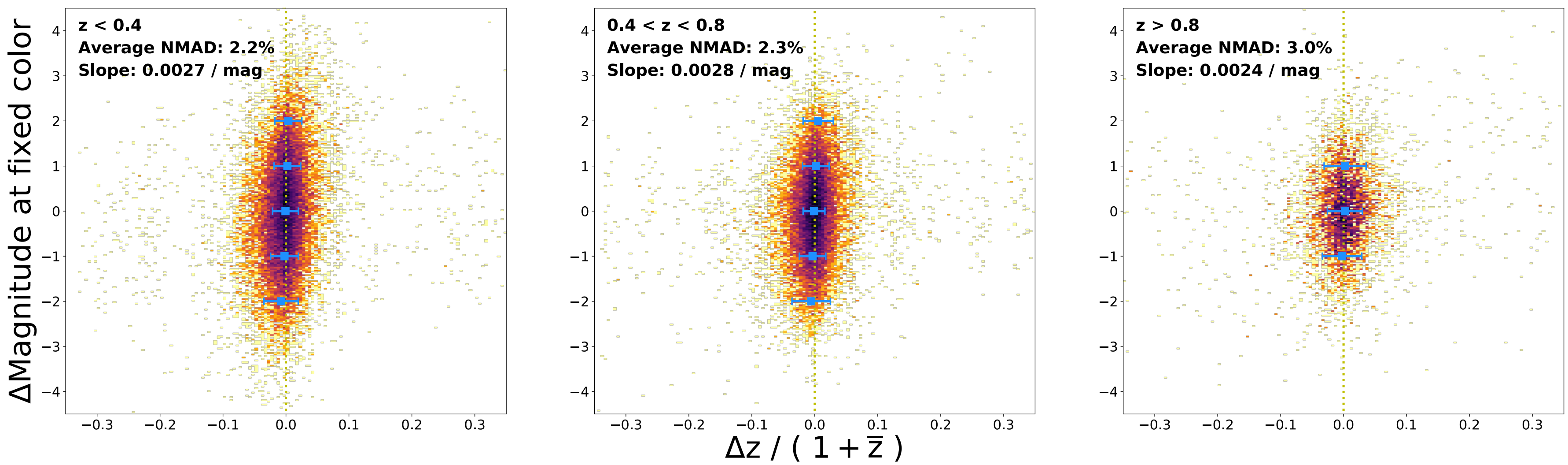


- Compare spec-zs of unique pairs of galaxies at fixed color (SOM position)
- Illustrates weak (and measurable) secondary dependence of redshift on magnitude *at fixed color* in the LSST+Euclid color space

Outlier fraction 4.1%, scatter 2.3%, bias of -0.2%

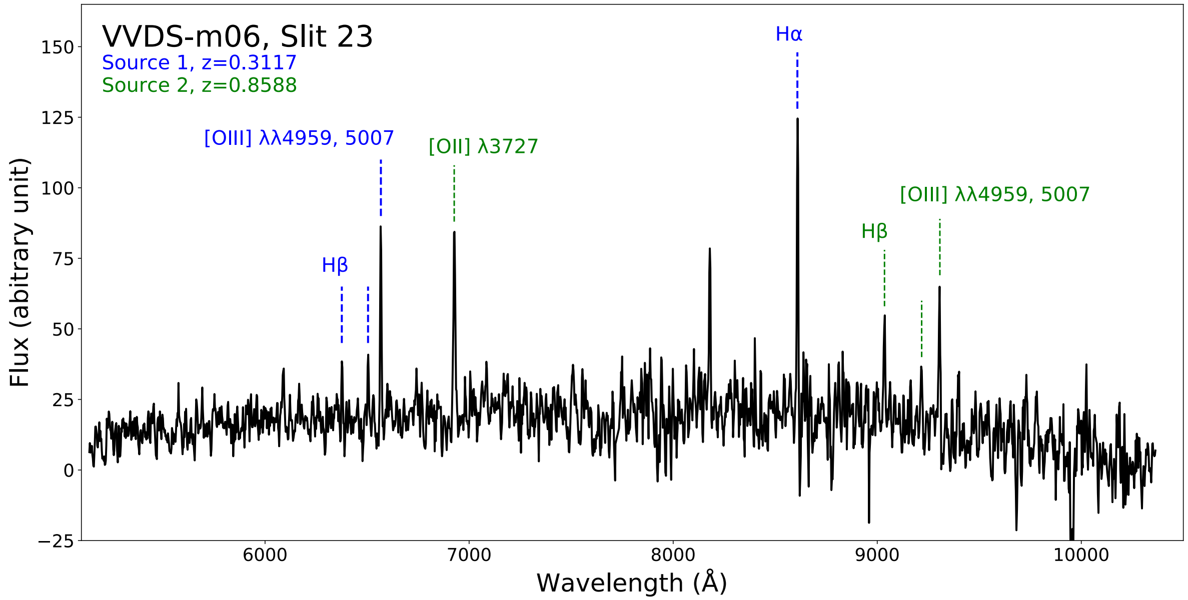
→ Method achieving unbiased performance

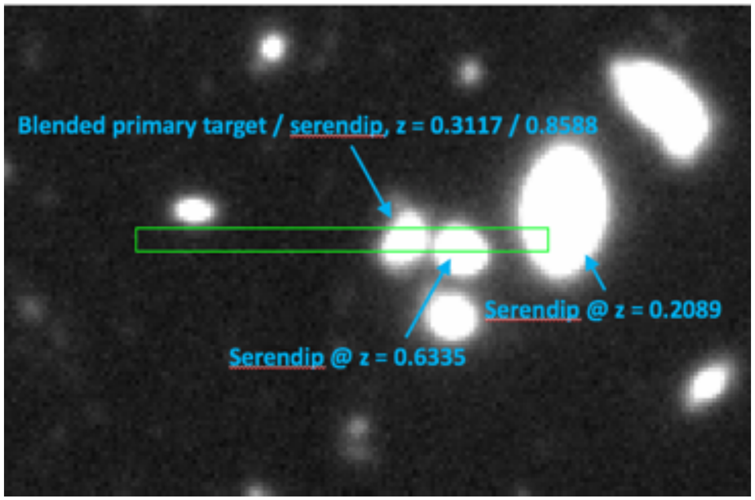
Remarkably stable relationship of dmag/dz at fixed color



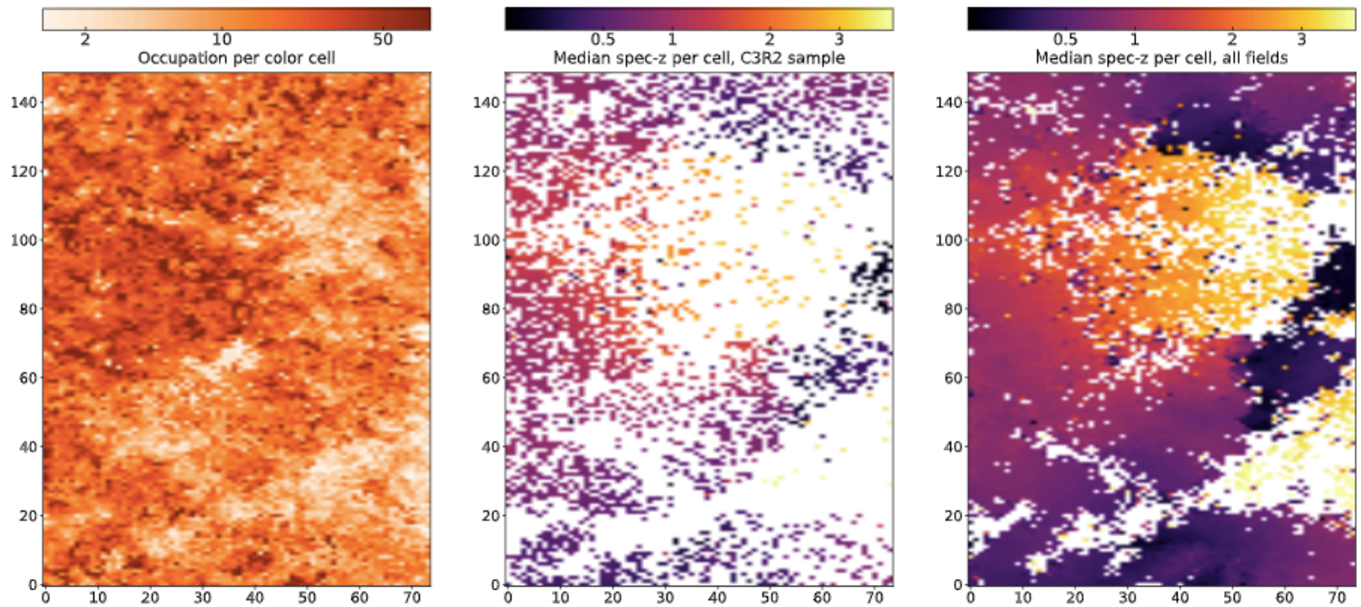
Hemmati et al. 2018

Ambiguous redshifts





Color coverage to Euclid depth



- Much of the galaxy color space explored to significant depth (>75% of cells; >85% of galaxies in cell with at least 1 specz, many cells with >>1 specz); some cells remain uncalibrated at present.
- C3R2-Keck alone covering >35% of the color space

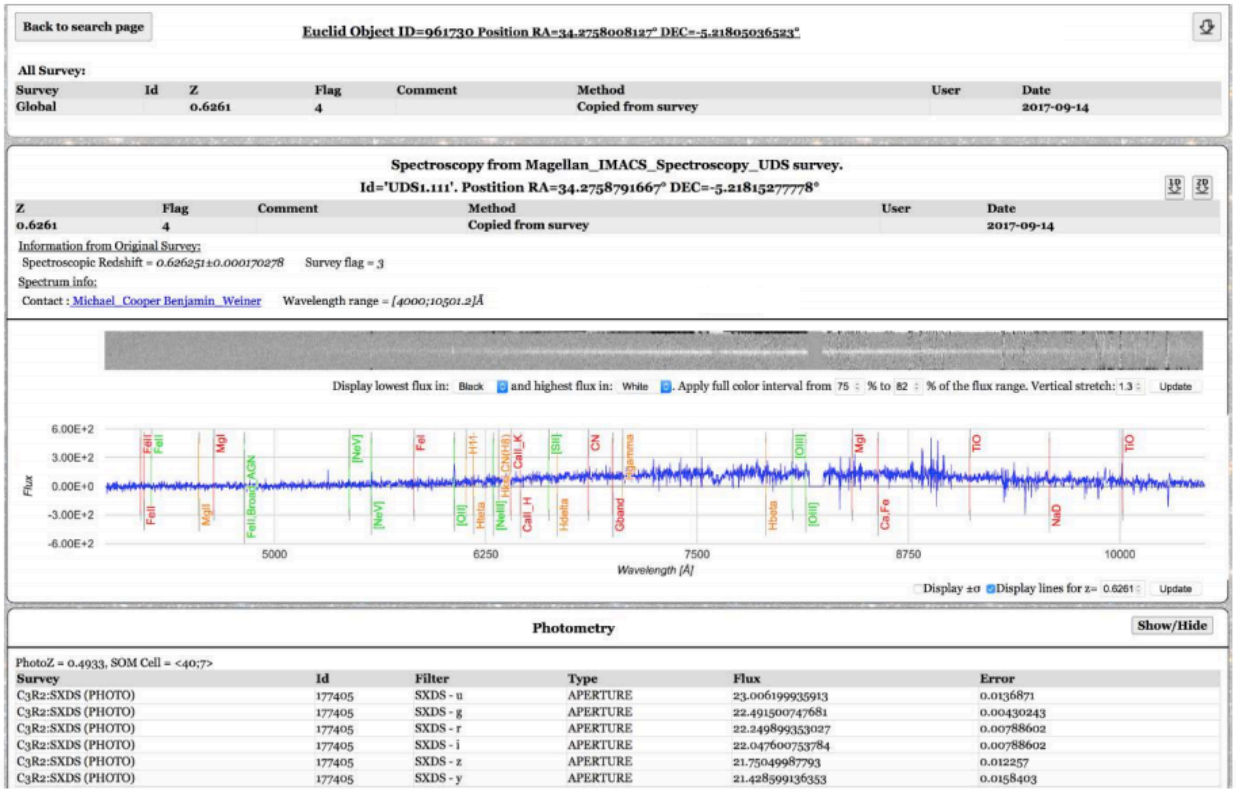
Spectroscopic Databases – challenges

- We will have hundreds of thousands of deep galaxy spectra in the 2020s, growing continuously
- Need a database that can easily ingest new spectroscopy from disparate sources (e.g., from grisms)
- Extremely careful vetting of spec-zs necessary for cosmology
- Machine learning-based redshifts may prove critical

All large cosmology surveys would benefit from a single high-quality database of all deep spectroscopy!

A start for Euclid in Geneva

The database – Case example / Read-only



Back to search page

Euclid Object ID=961730 Position RA=34.2758008127° DEC=-5.21805036523°

All Survey:

Survey	Id	Z	Flag	Comment	Method	User	Date
Global		0.6261	4		Copied from survey		2017-09-14

Spectroscopy from Magellan_IMACS_Spectroscopy_UDS survey.
Id='UDS1.111'. Postition RA=34.2758791667° DEC=-5.21815277778°

Z	Flag	Comment	Method	User	Date
0.6261	4		Copied from survey		2017-09-14

Information from Original Survey:
Spectroscopic Redshift = *0.626251±0.000170278* Survey flag = *3*
Spectrum info:
Contact : Michael_Cooper Benjamin_Weiner Wavelength range = *[4000;10501.2]Å*

Display lowest flux in: Black and highest flux in: White. Apply full color interval from 75 % to 82 % of the flux range. Vertical stretch: 1.3 Update

Display ±σ Display lines for z= 0.6261 Update

Photometry Show/Hide

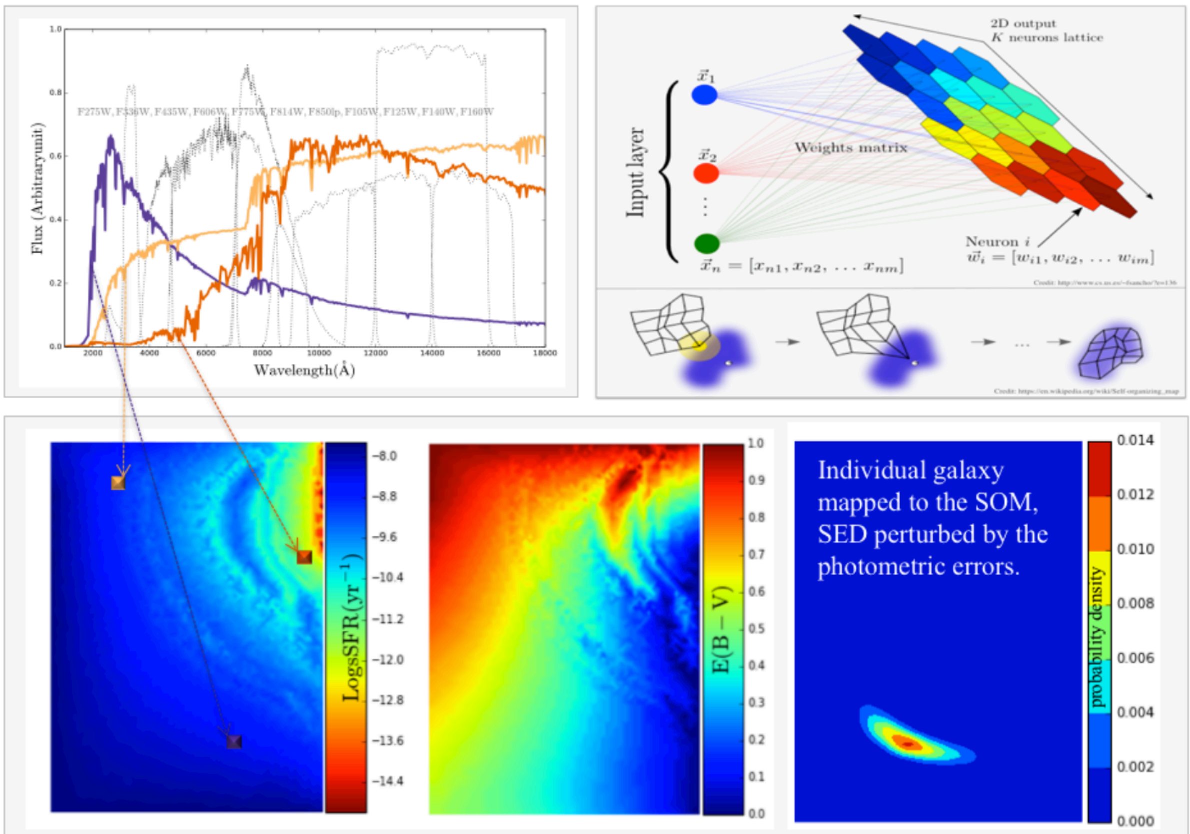
PhotoZ = 0.4933, SOM Cell = <40;7>

Survey	Id	Filter	Type	Flux	Error
C3R2:SXDS (PHOTO)	177405	SXDS - u	APERTURE	23.006199935913	0.0136871
C3R2:SXDS (PHOTO)	177405	SXDS - g	APERTURE	22.491500747681	0.00430243
C3R2:SXDS (PHOTO)	177405	SXDS - r	APERTURE	22.249899353027	0.00788602
C3R2:SXDS (PHOTO)	177405	SXDS - i	APERTURE	22.047600753784	0.00788602
C3R2:SXDS (PHOTO)	177405	SXDS - z	APERTURE	21.75049987793	0.012257
C3R2:SXDS (PHOTO)	177405	SXDS - y	APERTURE	21.428599136353	0.0158403

Galaxies are not unique

- The manifold of galaxy observables is finite, and large surveys like Euclid/WFIRST/LSST will measure essentially the same galaxy over and over
- We can measure the galaxy manifold really well with large surveys.
- Continuity constraints could then allow us to build a dynamic picture of galaxy growth
  - Individual galaxies can be thought of as moving along the manifold.
- What could we learn from this?

Models – can we match them to the data?



Measure the high-dimensional manifold. Then what?

- We have a well-defined target for simulations
- What if we find (as is common) that the simulations produce unphysical galaxies, or can't produce certain real galaxies?
- Is there a way to systematically search for the simulation parameters that produce the observed universe?
- What have we learned about galaxies at the end?
- How do we achieve the "Standard Model"?